Projekt

Cześć I

Grupy tydzień nieparzysty (nr 1, 2, 4, 6):
- Wprowadzenie: 14 marca 2025
- Termin wysłania Części I Projektu: 27 marca 2025
- Termin oddania: 28 marca 2025

Grypy tydzień parzysty (nr 3, 5, 7):
- Wprowadzenie: 21 marca 2025
- Termin wysłania Części I Projektu: 3 kwietnia 2025
- Termin oddania: 4 kwietnia 2025

Celem pierwszej części projektu jest zapoznanie się z językiem Python oraz narzędziami do analizy i wizualizacji danych. Studenci nauczą się, jak przygotować środowisko programistyczne, zaimplementować podstawowe operacje na zbiorach danych oraz przeprowadzić eksploracyjną analizę danych (EDA). W trakcie realizacji projektu uczestnicy zdobędą praktyczne umiejętności w zakresie wczytywania danych, obliczania statystyk opisowych, identyfikacji brakujących wartości oraz eksploracji rozkładu cech numerycznych i kategorialnych. Analiza danych będzie obejmować zarówno metody statystyczne, jak i graficzne, wykorzystując popularne biblioteki takie jak `pandas`, `matplotlib` i `seaborn`.
Projekt ma na celu rozwinięcie umiejętności analizy danych oraz krytycznego myślenia o ich strukturze i zależnościach. Wykorzystane techniki i narzędzia pozwolą studentom lepiej zrozumieć mechanizmy pracy z danymi, co będzie cenną umiejętnością w przyszłych projektach badawczych i zawodowych.

Część implementacyjna

Zakres prac na ocenę 3.0 z części I
1. Przygotowanie środowiska Python, w tym plików potrzebnych do automatycznego uruchomienia środowiska wirtualnego (np. `requirements.txt`).
2. Przygotowanie pliku `README` z opisem umożliwiającym uruchomienie projektu i poszczególnych skryptów (najlepiej w języku angielskim).
3. Zaimplementowanie metody (lub klasy) do wczytywania zbioru danych w języku Python.
4. Przygotowanie skryptu do obliczania i zapisywania wstępnych statystyk cech (możesz je zapisać w pliku CSV):
    - Dla cech numerycznych: średnia, mediana, wartość minimalna, maksymalna, odchylenie standardowe, 5-ty i 95-ty percentyl, liczba brakujących wartości w kolumnie.
    - Dla cech kategorialnych: liczba unikalnych klas, liczba brakujących wartości w kolumnie, proporcja klas.

Zakres prac na ocenę 3.5 z części I

1. Wykorzystanie boxplotów: https://seaborn.pydata.org/tutorial/categorical.html#boxplots
2. Wykorzystanie violinplotów: https://seaborn.pydata.org/tutorial/categorical.html#violinplots

Zakres prac na ocenę 4.0 z części I

1. Analiza i wizualizacja cech numerycznych za pomocą *error bars*: https://seaborn.pydata.org/tutorial/error_bars.html
2. Prezentacja histogramów dla cech numerycznych: https://seaborn.pydata.org/tutorial/distributions.html
3. Wykorzystanie histogramów warunkowanych (z parametrem `hue`): https://seaborn.pydata.org/tutorial/distributions.html#conditioning-on-other-variables

Zakres prac na ocenę 4.5 z części I

1. Przygotowanie heatmapy korelacji danych: https://seaborn.pydata.org/generated/seaborn.heatmap.html

Zakres prac na ocenę 5.0 z części I

1. Zaimplementowanie metody do analizy i wizualizacji korelacji liniowej między wartościami cech za pomocą regresji liniowej: https://seaborn.pydata.org/tutorial/regression.html

Zakres prac na ocenę 5.5 z części I

1. Przygotowanie wizualizacji danych z wykorzystaniem redukcji wymiarowości, np. PCA lub t-SNE.

Przygotuj spójny skrypt `data_analysis_main.py`, który umożliwi wygenerowanie i zapisanie na dysku wszystkich wyników i wizualizacji.

Uwaga. Zastanów się w jaki sposób może przeprowadzić analizę cech kategorialnych, skoro większość narzędzi jest dostosowana do cech numerycznych.

Cześć raportowa

- Raport może być przygotowany w dowolnej formie tekstowej, np. PDF, prezentacja.
- Celem raportu jest przedstawienie najciekawszych zależności znalezionych w analizowanym zbiorze danych, np. zależności między cechami, brak istotności niektórych zmiennych.
- W raporcie należy skupić się na zamieszczeniu wizualizacji wygenerowanych za pomocą napisanych skryptów, a także na analizie wyników i wnioskach.
- Obserwacje należy przedstawiać w zwięzłej formie, np. w kilku punktach pod każdym wykresem.
- Nie należy prezentować wszystkich cech w raporcie. Wybierz 3–5 cech, które mają największą wartość informacyjną.
- Raport powinien pokazywać zrozumienie wykorzystanych metod analizy danych.

Przykładowe pytania, na które można odpowiedzieć w raporcie (ale nie ograniczaj się tylko do nich):

- Czy w zbiorze występują silnie skorelowane zmienne?
- Jakie kategorie dominują w cechach kategorialnych?
- Czy cechy kategorialne, a szczególnie cecha docelowa (target), są zbalansowane?
- Czy wartości odstające mają istotny wpływ na dane?

Zbiory danych

Przy wyborze zbioru danych zwróć uwagę na jego wielkość – powinien zawierać co najmniej 2000 rekordów oraz przynajmniej 15 cech. Wybierz zbiór dostępny w łatwym do użycia formacie, np. CSV. Dane powinny zawierać zarówno cechy numeryczne, jak i kategorialne, aby umożliwić kompleksową analizę. Wszelkie odstępstwa od tych założeń należy skonsultować z prowadzącym.

Przykładowe zbiory danych, możliwe to wykorzystania w ramach kursu:

- Default of Credit Card Clients Dataset

Link: Kaggle https://www.kaggle.com/datasets/uciml/default-of-credit-card-clients-dataset
UC Irvine ML Repos: https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients

- FIFA 22 complete player dataset

Opis: 19k+ graczy, 100+ atrybutów pobranych z gry FIFA 2022
Link: Kaggle:
https://www.kaggle.com/datasets/stefanoleone992/fifa-22-complete-player-dataset

- Predict Students' Dropout and Academic Success

Link:
Kaggle:
https://www.kaggle.com/datasets/naveenkumar20bps1137/predict-students-dropout-and-academic-success
UC Irvine ML Repo:
https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success

- Diabetes 130-US Hospitals for Years 1999-2008

Uwaga. Dość duży zbiór
Link: UC Irvine ML Repo:
https://archive.ics.uci.edu/dataset/296/diabetes+130-us+hospitals+for+years+1999-2008

- Estimation of Obesity Levels Based On Eating Habits and Physical Condition

Link: UC Irvine ML Repo:
https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition

Wiecej zbiorów danych można znaleźć tutaj:
UC Irvine ML Repository: https://archive.ics.uci.edu/datasets
Kaggle: https://www.kaggle.com/datasets